# Optimal Administration of Dual Screening Tests for Detecting a Characteristic

by

Seymour Geisser* and Wesley Johnson
University of Minnesota and University of California, Davis

Technical Report No. 532
August 1989

## Summary

We consider the problem of deciding optimally whether a characteristic exists based on one or two screening tests. We discuss the relative merits of giving either one or two tests, including the order in which they might be given, as well as their costs. Operating in the Bayesian mode, we utilize posterior distributions for the accuracies of the tests, and the prevalence of the characteristic. Applications to detecting rare conditions, such as the AIDS virus, are discussed.

Key words: AIDS, FLV, Predictive value positive, prevalence, Bayesian approach, Dirichlet prior, sensitivity, specificity

*The work was sponsored in part by an NIH grant GMS-25271.

## 1. Introduction

With the introduction of multiple diagnostic tests for various conditions, including pre-clinical diseases, and the need that the public health of a population be carefully monitored so that appropriate informed actions can be taken when necessary, mass screening programs have been suggested. It is the purpose of this paper to set forth the requisite decision theoretic framework for such programs involving the estimation of the critical entities such as the prevalence, sensitivity, specificity, various predictive probabilities all combined with the costs of the tests and the utilities (or losses) for taking particular actions.

We assume a relevant population and a diagnostic test to ascertain the presence of a characteristic in a randomly drawn unit from the population. In the medical area, diagnostic tests are often used to attempt to confirm or indicate the preclinical or presymptomatic presence or near future onset of a disease. We first confine ourselves to a simple binary diagnostic test and develop the necessary tools for the full analysis of this situation. Then we consider two binary diagnostic tests given simultaneously or in sequence. Here the framework for the estimation of the attributes of the tests and their optimal administration for a mass screening program is considered. Costs of the tests and the losses associated with the various kinds of correct and incorrect diagnoses are also taken into account. In all that follows we assume that a diagnostic test is always capable of assigning every individual of a particular population as a positive or negative for the characteristic, even though this may not always be the case with real diagnostic tests. In particular the so-called gold standard test, that is the Western Blot test, for the presence of HIV antibodies can apparently result in four categories namely positive, probably positive, indeterminate and negative (Lundberg, 1988). Furthermore, ELISA tests often give equivocal results. We interpret equivocal results as positive.

Comparisons of two or more screening tests are made routinely in the medical and veterinary literature, for example see Burkhardt, Mertens, and Eggers, (1987); Sandstrom and Mortimer, (1987); Ghi et al., (1988); Hirsch, Searcy, and Bellamy, (1982); Jarrett, Golder, and Weijer, (1982); and Kurlander, Hill and Enterline, (1955). Most of these articles compare single tests with one another; Kurlander et al. (1955) consider combined tests. Data obtained in such

studies are generally not "ideal". An absolute gold standard often does not exist; while in other instances, not all individuals are subjected to the gold standard when it exists. Future experiments should obtain full information regarding the joint accuracies of multiple tests by confirming not only questionable test results, but also those results that are apparently, though not necessarily, unequivocal. Gastwirth (1987) discussed the problem of making inferences from screening data and gave many references to previous work. Geisser (1987) discussed the Bayesian approach with special emphasis on prediction problems. Gastwirth and Hammick (1989) considered pooling blood so that anonymity could be maintained while estimating prevalence. Johnson and Gastwirth (1988) and Gastwirth, Johnson and Reneau (1988) developed Bayesian methodology for making approximate statistical inferences about prevalence, sensitivity, specificity, etc. based on screening data. Gatsonis and Iyengar (1987) developed Bayesian methodology for screening data based on the approximations developed by Tierney and Kadane (1986).

We detail the setting for a single diagnostic test in section 2, and for two diagnostic tests in section 3. In section 4, we incorporate the possibility of differential costs for the various possible administrations of the two tests. We illustrate our procedures in section 5, and make some final remarks in section 6.

## 2. A Single Diagnostic Test

In order to introduce ideas we first consider a single diagnostic test. Define the following relevant quantities:

$$\pi = \Pr(C)$$

the probability that a randomly drawn individual from the population exhibiting characteristic C,

$$\eta = \Pr(T|C)$$

the probability that the test correctly diagnoses the absence of C, which is called the sensitivity,

$$\theta = \Pr(\bar{T}|\bar{C})$$

the probability that the test correctly diagnoses the absence of $\bar{C}$, which is called the specificity,

$$\psi = \Pr(C|T) = \pi\eta/(\pi\eta + (1-\pi)(1-\theta))$$

the probability that the characteristic is present given the test indicates its presence, which is called the predictive value positive (PVP), and

$$\bar{\psi} = \Pr(\bar{C}|\bar{T}) = (1-\pi)\theta/\{\pi(1-\eta) + (1-\pi)\theta\}$$

the probability that the characteristic is absent given the test indicates its absence, which is called the predictive value negative (PVN). When all of these entities are known, conceptually we need only introduce the costs, or utilities, and a few manipulations to determine an optimal decision procedure given an outcome on the diagnostic test.

Precise knowledge of $\pi$, $\eta$ and $\theta$ is often unavailable, or only partially so, in terms of samples from various subpopulations within the relevant population. Normally $\theta$ and $\eta$ will be unknown and their estimation would require independent random samples from two reference populations; the test is applied to n individuals or units known to have the characteristic, and also to $\bar{n}$ individuals known to be free of the characteristic. Assuming that r out of n yield T in the first sample and $\bar{r}$ out of $\bar{n}$ yield $\bar{T}$ in the second, we obtain the likelihood

$$L(\eta,\theta) \propto \eta^{r}(1-\eta)^{n-r}\theta^{\bar{r}}(1-\theta)^{\bar{n}-\bar{r}}.$$

If $\pi$ is unknown, and an independent sample of size v from the entire population is available, we have the additional likelihood contribution

$$L(\pi) \propto \pi^{t_c}(1-\pi)^{v-t_c}, \tag{2.1}$$

where $t_c$ represents the number of individuals having C. If this is unavailable, another independent sample of size s from the population could be obtained. In this sample we ascertain the number t that test T. Since
$\Pr(T) = \pi\eta + (1-\pi)(1-\theta)$, we obtain the joint likelihood

$$L(\eta,\theta,\pi) \propto \eta^{r}(1-\eta)^{n-r}\theta^{\bar{r}}(1-\theta)^{\bar{n}-\bar{r}}\{\pi\eta + (1-\pi)(1-\theta)\}^{t}\{\pi(1-\eta) + (1-\pi)\theta\}^{s-t}.$$

Alternatively, a random sample of size $n+\bar{n}$ from the population of interest could be obtained and each individual given the screening test as well as a "gold standard" confirmatory test. In this case, one finds that the likelihood is in exactly the same form as $L(\eta,\theta)L(\pi)$, thus we do not discuss it further.

Given a prior density, $p(\eta,\theta,\pi)$, and with $d = (r,n,\bar{r},\bar{n},t_c,v)$, the joint posterior density is obtained as

$$p(\eta,\theta,\pi|d) \propto p(\eta,\theta,\pi)L(\theta,\eta)L(\pi).$$

With $d = (r,n,\bar{r},\bar{n},t,s)$,

$$p(\eta,\theta,\pi|d) \propto p(\eta,\theta,\pi)L(\theta,\eta,\pi).$$

In both cases, the main quantities of interest are the predictive probabilities

$$\Pr(C|(+),d) = E(\psi|(+),d) = E(\pi\eta|d)/\{E(\pi\eta|d) + E((1-\pi)(1-\theta)|d)\}, \quad (2.2)$$

$$\Pr(\bar{C}|(-),d) = E(\bar{\psi}|(-),d) = E(\theta(1-\pi)|d)/\{E(\theta(1-\pi)|d) + E(\pi(1-\eta)|d)\}, \quad (2.3)$$

both obtained by simple application of Bayes Theorem. Gastwirth, Johnson and Reneau (1988) and Johnson and Gastwirth (1988) gave results in this setting where the prevalence was low and the accuracies were high.

Clearly, the predictive probability is the same for all exchangeable subjects. Hence, when a positive result is obtained, a decision regarding various alternative actions depends on the particular subject's costs, or losses. Assume that there are "a" actions and that for action i, the generic losses are $\ell_{iC}$ and $\ell_{i\bar{C}}$, the loss associated with taking action i when C is true and of taking action i when $\bar{C}$ is true, respectively; i=1,2,...,a. Then we need only minimize

$$E(L_i|(+)) = (\ell_{iC} - \ell_{i\bar{C}})\Pr(C|(+),d) + \ell_{i\bar{C}}$$

with respect to i, to obtain the optimal action. The predictive probability of C is the same for any individual that tests positive, but individuals may very well differ in the utilities they would assign. If expenses for each action, say $c_i$, are in the same units as the losses, then we would minimize $E(L_i|(+)) + c_i$. A similar analysis could hold for those that tested negative.

Another point of considerable importance is the potential variation in the prevalence of the characteristic amongst subpopulations at any given time. For example, with AIDS it is well known that there are substantial increases over the general population in the prevalence of the Human Immunodeficiency Virus among intravenous drug users, homosexuals, and hemophiliacs. In other situations there are population characteristics, which are neither lethal nor contagious, for which the prevalence may be more or less stable over time. However with highly contagious diseases this is obviously not so. Depending on the ease of transmission, time varying prevalence could be of critical importance if the screening is to be made over a protracted time period. We shall restrict ourselves here to situations where $\pi$ is considered to be largely stable.

We defer discussion of choice of prior and actual calculations to the end of the next section. Results for single tests can be obtained as special cases of results in section 3.

3. Two Tests

In this section, we allow for the availability of two distinct tests that may be administered either simultaneously or sequentially in a screening program. There are a number of different decision rules one can follow regarding how to proceed. If one were to administer two tests simultaneously, one could assert that the characteristic exists or pre-exists if both tests are positive or if either test is positive. A decision could be made to ignore one of the test results and to simply designate C or $\bar{C}$ according to a single test's result. One could administer the tests sequentially and allow for the possibility of administering one test first and then either stopping if the result were positive, or proceeding to a second test if the result were negative. The final decision would be to designate as C either a positive on the first test or a negative on the first and a positive on the second. Another sequential possibility is to give a single test and to designate $\bar{C}$ if the result is negative, otherwise give the second test and designate C only if both tests are positive.

We denote our single test results as $(T_i, \bar{T}_i)$ i=1,2 and introduce notation to indicate the above decision rules. There are eight possibilities that we will discuss. They are listed in Table 1.

Table 1

| Rule | Decision Rule (Assert C if) | Notation |
|---|---|---|
| 1 | Test 1 is positive | $T_1$ |
| 2 | Test 2 is positive | $T_2$ |
| 3 | Both tests positive (simultaneous tests) | $(T_1T_2)$ |
| 4 | Either test positive (simultaneous tests) | $(T_1 \cup T_2)$ |
| 5 | Both tests positive (sequential tests) | $T_1T_2$ |
| 6 | Both tests positive (sequential tests) | $T_2T_1$ |
| 7 | Either test positive (sequential tests) | $T_1 \cup \bar{T}_1T_2$ |
| 8 | Either test positive (sequential tests) | $T_2 \cup \bar{T}_2T_1$ |

Decision rule, #8 say, resulted in a positive result, hereafter denoted (+), if test #2 was administered first and was positive or if it was negative and then test #1 was subsequently positive. The eight omitted possibilities correspond to invalid tests; for example, assert C if Test 1 is negative.

There are expenses associated with the tests, and losses in making wrong

decisions. We defer discussion of expenses until section 4. Here, we define a function which reflects the various losses associated with correct and incorrect decisions, and in Section 4 we define a function which reflects the actual costs of administration associated with the decision rules.

We define our loss function in Table 2.

Table 2

| Decision Rule Outcome | True State C | True State $\bar{C}$ |
|---|---|---|
| + | $\ell_{11}$ | $\ell_{10}$ |
| - | $\ell_{01}$ | $\ell_{00}$ |

For example, the cost of a positive decision when C is present is $\ell_{11}$. The first subscript of $\ell_{ij}$ denotes the decision, positive (+) or negative (-), and the second denotes presence or absence of the characteristic.

When C denotes, say, the virus for AIDS, there are two points of view that one can take regarding this cost function; one from the perspective of society and the other from that of the individual. From a societal standpoint, it could be sensible to argue that

$$\ell_{11} \leq \ell_{00} < \ell_{10} \leq \ell_{01}.$$

In the examples section, we consider a situation where the cost associated with infecting an individual by transfusion of contaminated blood is assumed to be $\ell_{01} = \$10^6$. While there are costs associated with false positive results, they are expected to be much less than $\$10^6$. The individual involved would generally not be informed of a positive result unless it was confirmed by additional testing to be positive. Thus the loss $\ell_{10}$ would be expected to range from a minimum of \$15, depending upon how many confirmatory tests were equivocal before a definite decision could be made regarding the infection status. In the event that individuals were informed as to the results of the screening test, these costs could be much higher; the ordering could vary from one individual to another. Drug testing is a good example of a situation where the ordering would vary depending upon one's perspective. In general, it would only be reasonable to assume $\ell_{ii} < \ell_{kj}$ for all $k \neq j$.

A generalization of our notation from section 2 regarding the conditional probabilities for the various outcomes of the two tests is introduced in Table 3.

Table 3

| C | $T_2$ | $\bar{T}_2$ |
|---|---|---|
| $T_1$ | $\eta_{11}$ | $\eta_{10}$ |
| $\bar{T}_1$ | $\eta_{01}$ | $\eta_{00}$ |

| $\bar{C}$ | $T_2$ | $\bar{T}_2$ |
|---|---|---|
| $T_1$ | $\theta_{11}$ | $\theta_{10}$ |
| $\bar{T}_1$ | $\theta_{01}$ | $\theta_{00}$ |

Thus $Pr(T_1,T_2|C) = \eta_{11}$, $Pr(\bar{T}_1,T_2|\bar{C}) = \theta_{01}$, etc. We define $Pr(C) = \pi$ as before, and the predictive values positive and negative for decision rule i as $Pr(C|+) = \psi_i$ and $Pr(\bar{C}|-) = \bar{\psi}_i$, for i=1,2,...,8. For each of the eight joint tests in Table 1, define the sensitivity and specificity as

$$\eta_i = Pr((+)|C), \ \theta_i = Pr((-)|\bar{C}) \qquad i=1,\ldots,8.$$

For decision rule $R_i$, the expected loss is

$$E(L|R_i) = \pi\eta_i[\ell_{11}-\ell_{01}] + (1-\pi)\theta_i[\ell_{00}-\ell_{10}] + \pi\ell_{01} + (1-\pi)\ell_{10}. \qquad (3.1)$$

The values of $Pr((+)|C)$ and $Pr((-)|\bar{C})$ for each of the first 4 decision rules are presented in Table 4.

Table 4

| Rule # | $Pr(+\|C)$ (sensitivity) | $Pr(-\|\bar{C})$ (specificity) |
|---|---|---|
| 1 | $\eta_1 = \eta_{11} + \eta_{10}$ | $\theta_1 = \theta_{00} + \theta_{01}$ |
| 2 | $\eta_2 = \eta_{11} + \eta_{01}$ | $\theta_2 = \theta_{00} + \theta_{10}$ |
| 3 | $\eta_3 = \eta_{11}$ | $\theta_3 = \theta_{00} + \theta_{01}\ \theta_{10}$ |
| 4 | $\eta_4 = \eta_{11} + \eta_{10} + \eta_{01}$ | $\theta_4 = \theta_{00}$ |

Note from (3.1) that, irrespective of prevalence and losses resulting from either false positive or negative results, if $\eta_{10} = \eta_{01} = 0$, $(T_1T_2)$ is optimal. If $\theta_{01} = \theta_{10} = 0$, $(T_1 \cup T_2)$ is optimal. If $\eta_{10} = \theta_{01} = 0$, $T_2$ is optimal, and if $\eta_{01} = \theta_{10} = 0$, $T_1$ is optimal. Furthermore, it follows easily that the values for rules #5 and #6 are equivalent to those for rule #3, and similarly rules #7 and #8 are equivalent to rule #4.

For a mass screening program, such as contemplated for certain diseases, it would be desirable to use the optimal rule. Now rule R will be preferred to R* if the expected loss under R is less than that under R*. Denote this as $R > R^*$. It is straightforward to show that if $k = (\ell_{10}-\ell_{00})/(\ell_{10}-\ell_{00}+\ell_{01}-\ell_{11})$, then

$$\begin{aligned} T_2 > (T_1 \cup T_2) &\Leftrightarrow (T_1 T_2) > T_1 \Leftrightarrow \Pr(C|T_1\bar{T}_2) < k \\ T_1 > (T_1 \cup T_2) &\Leftrightarrow (T_1 T_2) > T_2 \Leftrightarrow \Pr(C|\bar{T}_1 T_2) < k \\ (T_1 T_2) > (T_1 \cup T_2) &\Leftrightarrow \Pr(C|\bar{T}_1 T_2 \cup T_1 \bar{T}_2) < k \end{aligned} \tag{3.2}$$

$$T_2 > T_1 \Leftrightarrow \frac{\Pr(C, T_1\bar{T}_2) - \Pr(C, \bar{T}_1 T_2)}{\Pr(T_1\bar{T}_2) - \Pr(\bar{T}_1 T_2)} < k$$

$$\Pr(C|T_1\bar{T}_2) < \Pr(C|T_1\bar{T}_2 \cup \bar{T}_1 T_2) < \Pr(C|\bar{T}_1 T_2) \Leftrightarrow \Pr(C|T_1\bar{T}_2) < \Pr(C|\bar{T}_1 T_2).$$

Note that $k = 1/2$ when $\ell_{10}-\ell_{00} = \ell_{01}-\ell_{11}$ or more particularly if $\ell_{00} = \ell_{11}$ and $\ell_{10} = \ell_{01}$. Suppose that $\Pr(C|T_1\bar{T}_2) < k < \Pr(C|\bar{T}_1 T_2)$. Then it follows from (3.2) that

$$T_2 > (T_1 \cup T_2) > T_1; \text{ and } T_2 > (T_1 T_2) > T_1,$$

thus $T_2$ would be the optimal procedure if the expense of testing were irrelevant. In fact, it is also the case that if $T_2$ is optimal, then the above condition must hold, due to (3.2), so the condition is necessary and sufficient. The following necessary and sufficient conditions are shown using (3.2):

$$\begin{aligned} T_1 \text{ is optimal} &\Leftrightarrow \Pr(C|\bar{T}_1 T_2) < k < \Pr(C|T_1\bar{T}_2) \\ T_2 \text{ is optimal} &\Leftrightarrow \Pr(C|T_1\bar{T}_2) < k < \Pr(C|\bar{T}_1 T_2) \\ (T_1 T_2) \text{ is optimal} &\Leftrightarrow \Pr(C|T_1\bar{T}_2) < k \text{ and } \Pr(C|\bar{T}_1 T_2) < k \\ (T_1 \cup T_2) \text{ is optimal} &\Leftrightarrow \Pr(C|T_1\bar{T}_2) > k \text{ and } \Pr(C|\bar{T}_1 T_2) > k. \end{aligned} \tag{3.3}$$

These conditions translate easily into equivalent statements regarding the parameters $\{\eta_{ij}, \theta_{ij}\}$ and $\pi$.

With the values in Table 3 fixed and known, it is possible to partition the space of $(k,\pi)$ values into regions for which the various rules 1-4 are optimal. Define $k^* = k/(1-k)$, $a = \theta_{01}/\eta_{01}$, $b = \theta_{10}/\eta_{10}$, $a^* = \theta_{01}/(\theta_{01}+\eta_{01})$, $b^* = \theta_{10}/(\theta_{10}+\eta_{10})$. Then

$T_1$ is optimal $\Leftrightarrow \pi < ak^*/(1+ak^*),\ \pi > bk^*/(1+bk^*)$

$T_2$ is optimal $\Leftrightarrow \pi < bk^*/(1+bk^*),\ \pi > ak^*/(1+ak^*)$

$(T_1T_2)$ is optimal $\Leftrightarrow \pi < \min\{ak^*/(1+ak^*),\ bk^*/(1+bk^*)\}$

$(T_1 \cup T_2)$ is optimal $\Leftrightarrow \pi > \max\{ak^*/(1+ak^*),\ bk^*/(1+bk^*)\}$

If $a < b$, the regions of optimality are given in Figure 1. If $a > b$, then the middle region corresponds to values of $(k^*,\pi)$ for which $T_1$ is optimal; $T_2$ would thus never be optimal when $a > b$.

Now assume the parameters are unknown and that prior information is available for $(\{\eta_{ij}\}, \{\theta_{ij}\}, \pi)$ in the form of independent Dirichlet distributions; namely

$$p(\{\eta_{ij}\}) \propto \prod_{i,j} \eta_{ij}^{\alpha_{ij}-1} \qquad \eta_{ij} \geq 0,\ \sum_{i,j} \eta_{ij} = 1,\ \sum_{i,j} \alpha_{ij} = \alpha,$$

$$p(\{\theta_{ij}\}) \propto \prod_{i,j} \theta_{ij}^{\bar{\alpha}_{ij}-1} \qquad \theta_{ij} \geq 0,\ \sum_{i,j} \theta_{ij} = 1,\ \sum_{i,j} \bar{\alpha}_{ij} = \bar{\alpha}, \tag{3.5}$$

$$p(\pi) \propto \pi^{\gamma-1}(1-\pi)^{\bar{\gamma}-1} \qquad 0 \leq \pi \leq 1.$$

Assume "training data" is available, where individuals known to have C and individuals known to be $\bar{C}$ are tested and a decision made as to (+) or (-) in each case. The likelihood corresponding to such data is

$$L(\{\eta_{ij}\},\{\theta_{ij}\}) = \prod_{i,j} \eta_{ij}^{r_{ij}} \theta_{ij}^{\bar{r}_{ij}};\ \sum_{i,j} r_{ij} = n,\ \sum_{i,j} \bar{r}_{ij} = \bar{n}.$$

Data for $\pi$ can either be direct or indirect, as in the one test case discussed in section 2. Direct data for $\pi$ is assumed to arise from a binomial sample; the corresponding likelihood is the same as in (2.1). Indirect data for a $\pi$ is obtained from screening data on individuals whose status with respect to C is unknown. This type of data is especially appropriate for small $\pi$, since direct estimates of $\pi$ in such cases would often be prohibitively expensive and/or impossible to obtain. Gastwirth et al. (1988) and Johnson and Gastwirth (1988) detail Bayesian methods for estimating low prevalence from screening data based on a single test. We assume indirect data for $\pi$ is in the form of a screening test from the general population under scrutiny. Two types of screening data

based on joint testing are

$S_1$: $s_1$ individuals are randomly selected and tested with test #1 and $s_2$ individuals are randomly selected and tested with test #2

$S_2$: s individuals are randomly selected and tested with both tests #1 and #2.

The likelihoods for $S_1$ and $S_2$ are:

$$L_1(\{\eta_{ij}\},\{\theta_{ij}\},\pi) = \prod_{i=1}^{2} \{\pi\eta_i + (1-\pi)(1-\theta_i)\}^{t_i} \{\pi(1-\eta_i) + (1-\pi)\theta_i\}^{s_i - t_i}$$

$$L_2(\{\eta_{ij}\},\{\theta_{ij}\},\pi) = \prod_{i,j} \{\pi\eta_{ij} + (1-\pi)\theta_{ij}\}^{t_{ij}}.$$

Screening data could also be obtained from the other sampling schemes.

For the above sampling designs, the total likelihoods are $L(\pi)L(\{\eta_{ij}\},\{\theta_{ij}\})$ and $L_i(\{\eta_{ij}\},\{\theta_{ij}\},\pi)L(\{\eta_{ij}\},\{\theta_{ij}\})$, i=1,2. With direct information for $\pi$, the posteriors for $\{\eta_{ij}\},\{\theta_{ij}\}$ and $\pi$ are independent and Dirichlet. Writing "d" for data,

$$\{\eta_{ij}\}|d \sim D(\{r_{ij} + \alpha_{ij}\}),\ \{\theta_{ij}\}|d \sim D(\{\bar{r}_{ij} + \bar{\alpha}_{ij}\})$$
$$\pi|d \sim \text{Beta}(\gamma + t_c,\ \bar{\gamma} + v - t_c). \tag{3.6}$$

Posterior means and variances and the sensitivities and specificities of all decision rules are easily obtained under this setup. For example, the sensitivity for the test $T_1$ is estimated as

$$E(\eta_1|d) = (r_{11} + r_{10} + \alpha_{11} + \alpha_{10})/(n + \alpha);$$

the prevalence estimate is

$$E(\pi|d) = (\gamma + t_c)/\bar{\gamma} + v).$$

Joint posteriors under screening data $S_1$ and $S_2$ are easily obtained, but are computationally burdensome. Simplifications in the high accuracy and low prevalence case (Johnson and Gastwirth, 1988) are possible, but will be reported elsewhere. For the remainder of this paper, we assume that direct information is available for $\pi$, and thus the posterior distribution of all parameters is given by (3.6). In those cases where current data are unavailable, one's prior knowledge of $\pi$ in conjunction with data for the $\eta_{ij}$'s and $\theta_{ij}$'s may be sufficiently precise to warrant an analysis without current data for $\pi$.

The predictive probability of C for an individual that has just tested

positive according to decision rule i is calculated as $Pr(C|(+),d) = E(\psi_i|(+),d)$. Upon application of Bayes theorem, this can also be expressed as in (2.2) with $(\eta_i,\theta_i)$ substituted for $(\eta,\theta)$. Similarly, $Pr(\bar{C}|(-),d) = E(\bar{\psi}_i|(-),d)$. Under our sampling scheme for $\pi$, all expectations of products result in products of expectations due to independence. For example,

$$Pr(C|(+),d) = \frac{E(\pi|d)E(\eta_i|d)}{E(\pi|d)E(\eta_i|d) + E(1-\pi|d)E(1-\theta_i|d)} .$$

From (3.1) the expected loss, given rule i, is

$$E(L|R_i,d) = E(\pi\eta_i|d)(\ell_{11}-\ell_{01}) + E((1-\pi)\theta_i|d)(\ell_{00}-\ell_{10}) + E(\pi|d)\ell_{01} + E(1-\pi|d)\ell_{10}. \tag{3.7}$$

Upon substitution of particular values for the sensitivity and specificity for different rules, it follows from (3.7) that (3.2) and (3.3) must hold with all conditional probabilities replaced with corresponding predictive probabilities. For example, replace $Pr(C|T_1\bar{T}_2)$ with $Pr(C|T_1\bar{T}_2,d) = E(\pi\eta_{10}|d)/E(\pi\eta_{10} + (1-\pi)\theta_{10}|d)$. Furthermore replace $Pr(C,T_1\bar{T}_2)$, $Pr(T_1\bar{T}_2)$ etc. with their posterior expectations.

## 4. Costs of Administration

In any large scale screening program costs of administering the tests will be of considerable concern. For example, with regard to AIDS, the ELISA test is an order of magnitude less expensive than the Western Blot. Also, there may be a differential in the cost of a simultaneous administration of both tests in contrast to their sequential administration. Among other things this may result from having to store the sample until the result from the first test is obtained, or from asking a testee to return for testing.

Once all of the actual testing costs are carefully ascertained, their incorporation into a complete decision analysis can be made without much difficulty. The major problems are in assessing, in some reasonable way, the original losses on a comparable monetary scale with the actual expense of testing. Another decision analytic consideration would be for a situation where a program was limited to a fixed amount of funding because of competing concerns. In this case, procedures that optimize certain well-defined benefits would be of critical interest.

Let $K_i$ be the cost of administering test i alone, and let $K_{ij}$ be the cost of administering test i followed by administering test j. Let $K_{(12)}$ be the cost of administering both tests simultaneously. Clearly, it is reasonable to assume that $K_{ij} \geq K_{(12)} \geq \max(K_1, K_2)$. In testing for a disease, for example, there may be storage and retrieval costs for sequential sampling. The expected costs for the eight decision rules are given in Table 5.

Table 5

| Decision Rule | E(Cost) |
|---|---|
| 1 | $K_1$ |
| 2 | $K_2$ |
| 3 | $K_{(12)}$ |
| 4 | $K_{(12)}$ |
| 5 | $K_1 + (K_{12}-K_1)\Pr(T_1)$ |
| 6 | $K_2 + (K_{21}-K_2)\Pr(T_2)$ |
| 7 | $K_1 + (K_{12}-K_1)\Pr(\bar{T}_1)$ |
| 8 | $K_2 + (K_{21}-K_2)\Pr(\bar{T}_2)$ |

The above probabilities are conditional on the parameters. If they were unknown, then they would be conditional on d.

Define the total expected loss for decision rule i to be the sum of the appropriate expected loss from (3.1) or (3.7) and the appropriate expected cost from Table 5. The total expected loss for rule 5 is greater than that for rule 3 if and only if E(Cost) for rule 5 is greater than that for rule 3. The same holds true for rule 6 compared with rule 3. Rule 7 is preferable to rule 4 if E(Cost) for rule 7 is less than E(Cost) for rule 4, and the same holds true for rule 8 compared with rule 4. The decision as to whether to consider sequential tests versus simultaneous tests is based purely on costs. Thus if $K_{12} = K_{21} = K_{(12)}$ or more particularly if $K_{ij} = K_{(12)} = K_1+K_2$, then rules 5 and 6 will be preferable to rule 3 and rules 7 and 8 will be preferable to rule 4.

## 5. Illustrations

In this section we consider three data sets; two concerning AIDS and one concerning feline leukemia virus (FLV).

The first two data sets are abstracted from studies that were designed to compare different tests for detecting antibodies to the AIDS virus (Burkhardt et

al., (1987); Sandstrom et al., (1985)). Neither study contained information for $\pi$. In both of these studies, we utilize data for the accuracies in conjunction with independent information for $\pi$ from the population of blood donors in Canada (Nusbacher et al., (1986)).

For both of the AIDS studies, individual serum specimens were tested with two commercial preparations. For our purposes, weak positives and negatives were considered positive. In both studies, specimens that were weakly positive or negative on either test, or that resulted in discrepant results {((+)(-)) or ((-),(+)} were retested in a variety of ways, including the Western Blot test, until an ultimate determination as to the true status of the specimen was ascertained. In one of the studies, a subsample of {(+)(+)} and {(-)(-)} samples were given confirmatory tests and one discrepancy was noted. From these studies, we have abstracted 2×2×2 tables of counts in the form of Table 3. It will be assumed that these are the same tables that would have arisen if all specimens were given confirmatory tests. Since there are possible errors in our abstraction of the data, the analyses are only illustrative.

The third data set is assumed to have been obtained from blood samples taken from a random sample of cats (Jarrett et al., 1982). Three tests were performed on each specimen; one of the tests is assumed to be superior to the other two and will be considered to be a "gold standard". Thus, these data are assumed to contain information for $\pi$ as well as for the accuracies.

Assuming a generic measurement scale for the moment, let $\ell_{00} = \ell_{11} = 0$ and $\ell_{01} = 1 \geq \ell_{10} = q > 0$ for all those studies. This reflects a societal attitude that it is worse to infuse contaminated blood than it is to falsely identify a healthy donor in the AIDS examples; and that it is worse to misinform an individual cat owner that his/her cat is FLV free when it is infected than it is to misdiagnose the cat as having FLV when it doesn't. With these assumptions, $k = q/(1+q)$ and $k^* = q$.

For the priors, we assume $\alpha_{11} = 3.9 = \bar{\alpha}_{00}$, $\alpha_{10} = \alpha_{01} = .5 = \bar{\alpha}_{10} = \bar{\alpha}_{01}$, and $\alpha_{00} = .1 = \bar{\alpha}_{11}$ for all examples. Thus, our prior for $\{\eta_{ij}\}$ has a weight, $\alpha_{11}+\alpha_{10}+\alpha_{01}+\alpha_{00} = 5$, which is equivalent to a sample of size $n = 5$ for the data. The same statement holds for $\{\theta_{ij}\}$. Furthermore, our prior reflects the belief that simultaneous correct screening results are quite likely (prior probability $= 3.9/5 = .78$) and that incorrect results are considerably less likely, but still plausible.

Information for $\pi$ is the same for both AIDS studies. We utilize Canadian

data obtained by Nusbacher et al. (1986) and analyzed by Gastwirth, Johnson and Reneau (1988). The Canadians administered a questionnaire which was designed to encourage "high risk" individuals to donate their blood anonymously for research purposes rather than for possible transfusion. Among 94,946 individuals whose blood was not utilized for research purposes, 405 samples tested positive on an Abbott ELISA screening test, and among those, 14 were confirmed positive by Western Blot. If we assume the remaining 94,091 negative samples were actually uncontaminated, then these data could be treated as a Bin (94496,$\pi$) sample. The prevalence $\pi$ is the proportional contaminated samples in the Canadian blood that is available for transfusion. With an improper B(1,0) prior for $\pi$, $E(\pi|d)$ = .000148. Alternatively, one could simply assume a beta prior with mean .000148 and an appropriate standard deviation. For the FLV data, we assume a B(1,0) prior for $\pi$ since data is available for $\pi$ in the experiment.

Example 1: Burkhardt, Mertens, and Eggers (1987) sampled 503 individuals and tested each serum specimen with an Abbott ELISA test and a DuPont ELISA test. The following data on accuracies were abstracted from their study.

Table 6

| C | | Abbott + | Abbott - |
|---|---|---|---|
| DuPont | + | 92 | 0 |
| | - | 1 | 0 |

| $\bar{C}$ | | Abbott + | Abbott - |
|---|---|---|---|
| DuPont | + | 8 | 9 |
| | - | 23 | 370 |

We do not use information for $\pi$ or q as yet. Utilizing the data displayed in Table 6, we obtain independent Dirichlet posteriors for $(\{\eta_{ij}\}\{\theta_{ij}\},\pi)$.

The conditional predictive probabilities of positive and negative results for rules 1-4 are listed in Table 7.

Table 7

| Rule | $Pr(+\|C,d)$ (sensitivity) | $Pr(-\|\bar{C},d)$ (specificity) |
|---|---|---|
| DuPont | .984 | .958 |
| Abbott | .994 | .924 |
| (DuPont∩Abbott) | .979 | .981 |
| (DuPont∪Abbott) | .999 | .901 |

These probabilities are calculated as $Pr(+|C,d) = E(\eta|d)$, $Pr(-|\bar{C},d) = E(\theta|d)$.

In order to determine the optimality regions for rules 1-4, we obtain

$a = E(\theta_{01}|d)/E(\eta_{01}|d) = 3.70$ and $b = E(\theta_{10}|d)/E(\eta_{10}|d) = 4.50$. Optimality is determined according to (3.4) and is viewed pictorially by consideration of Figure 1 with $a^* = .79$, $b^* = .82$, $k^* = q$, and with $E(\pi|d)$ substituted for $\pi$. The DuPont ELISA alone would not be optimal for any q and regardless of the value of $E(\pi|d)$.

If the posterior for $\pi$ has mean .000148, as discussed above for the Canadian transfusion pool, it remains to consider q before making a decision regarding the choice among rules 1-4, without regard to costs of administering the tests. Suppose that the value $\$10^6$ is attached to the life of an individual that is transfused with contaminated blood. In this instance, the rule (DuPont $\cap$ Abbott) is preferable if $q > \$40$, the rule (DuPont $\cup$ Abbott) is preferable if $q < \$33$, and the rule (Abbott) is preferable if $\$33 < q < \$40$.

We now consider the expenses associated with the administration of these tests. For simplicity, assume $K_1 = K_2 = K$, $K_{(12)} = K_{12} = K_{21} = 2K$. Then equating $T_1$ = Dupont and $T_2$ = Abbott we obtain

$$E(\text{Cost}|R_1) = E(\text{Cost}|R_2) = K,\ E(\text{Cost}|R_3) = E(\text{Cost}|R_4) = 2K$$
$$E(\text{Cost}|R_5) = K\{1 + \Pr(T_1|d)\},\ E(\text{Cost}|R_6) = K\{1 + \Pr(T_2|d)\},$$
$$E(\text{Cost}|R_7) = K\{1 + \Pr(\bar{T}_1|d)\},\ E(\text{Cost}|R_8) = K\{1 + \Pr(\bar{T}_2|d)\},$$

where

$$\Pr(T_1|d) = .941\ E(\pi|d) + .042,\ \Pr(T_2|d) = .918\ E(\pi|d) + .076.$$

Thus for the situation above with $E(\pi|d) = .000148$,

$$E(\text{Cost}|R_5) \doteq 1.042K,\ E(\text{Cost}|R_6) \doteq 1.076K$$
$$E(\text{Cost}|R_7) \doteq 1.958K,\ E(\text{Cost}|R_8) \doteq 1.924K.$$

Suppose $K = \$1$ and $q = \$25$. Without regard to the cost of administering a test, $T_1 \cup T_2$ is optimal. The total expected losses, per individual, for the eight rules are \$4.42, \$3.79, \$5.58, \$4.62, \$4.46, \$3.87, \$5.54, and \$4.54, respectively. Thus $T_2$ would be optimal overall. If on the other hand $K = \$5$ and $q = \$50$, then $(T_1 \cap T_2)$ is optimal without considering test expense. The total expected costs for the eight rules are \$11.57, \$13.49, \$15.01, \$20.05, \$10.22, \$10.39, \$19.84, and \$19.67 respectively. Thus the sequential rule $T_1T_2$ would be optimal. The values of PVP and 1-PVN for the rules are given in Table 8.

Table 8

| Rule | $Pr(C\|(+),d) = PVP$ | $Pr(C\|(-),d) = 1\text{-}PVN$ |
|---|---|---|
| $T_1$ | .0035 | $2.4 \times 10^{-6}$ |
| $T_2$ | .0019 | $.8 \times 10^{-6}$ |
| $(T_1 \cap T_2)$ | .0076 | $3.1 \times 10^{-6}$ |
| $(T_1 \cup T_2)$ | .0015 | $.2 \times 10^{-6}$ |

Example 2: These data are abstracted from Sandstrom et al. (1985). A total of 225 sera were tested by indirect immunofluorescence assay (IFA) and with an ELISA test. The data is given in Table 9.

Table 9

| | C | ELISA + | ELISA - | | $\bar{C}$ | ELISA + | ELISA - |
|---|---|---|---|---|---|---|---|
| IFA | + | 139 | 4 | IFA | + | 7 | 0 |
| | - | 1 | 0 | | - | 23 | 51 |

Since a = 27.1, b = .19, a* = .96 and b* = .16, the ELISA is never optimal, and the IFA is optimal for far more values of $(q, E(\pi|d))$ than in the previous example. With $E(\pi|d) = .000148$, and with a $\$10^6$ cost of a false negative where $T_1$ = IFA and $T_2$ = ELISA, $T_1 \cap T_2$ is optimal for $q \geq \$778.95$, $T_1$ is optimal for $\$5.46 < q < \$778.95$ and $T_1 \cup T_2$ is optimal for $q \leq \$5.46$. Expenses of tests are not considered here. The PVP and 1-PVN for the rules are listed in Table 10.

Table 10

| Rule | PVP | 1-PVN |
|---|---|---|
| $T_1$ | .0017 | $1.7 \times 10^{-6}$ |
| $T_2$ | .0004 | $7.1 \times 10^{-6}$ |
| $T_1 \cap T_2$ | .0017 | $6.6 \times 10^{-6}$ |
| $T_1 \cup T_2$ | .0004 | $.2 \times 10^{-6}$ |

Example 3: These data were taken directly from Jarrett et al. (1982). In this study, 412 blood samples were taken from cats. Three tests for feline leukemia virus (FLV) were applied to each sample, namely the ELISA test, immunofluorescence (IFA) and virus isolation (VI). We assume VI is a "gold standard" for the purposes of our discussion, which it would evidently be under

ideal circumstances (Jarrett et al. 1982). The results were obtained from Table 3 of Jarrett et al. (1982) and are displayed in Table 11.

Table 11

| C | | ELISA + | ELISA - |
|---|---|---|---|
| IFA | + | 47 | 0 |
| | - | 2 | 0 |

| $\bar{C}$ | | ELISA + | ELISA - |
|---|---|---|---|
| IFA | + | 3 | 1 |
| | - | 20 | 339 |

Since a = 1.20, b = .44, a* = .55 and b* = .31, the ELISA alone is never optimal. Refer to Figure 1 to visualize the regions of optimality for the other tests. If the data from Table 11 were obtained as a random sample of size 412 from some meaningful population, like the population of all household cats, for example, we would obtain the likelihood contribution for $\pi$

$$L(\pi) \propto \pi^{49}(1-\pi)^{363}.$$

Assuming a B(1,0) prior for $\pi$, we obtain a B(50, 363) posterior and consequently $E(\pi|d) = .12$. Assuming a loss of \$100 for a false negative result, (IFA $\cup$ Elisa) is optimal if the cost of a false positive result q < \$11.36, IFA is optimal if \$11.36 < q < \$31.00, and (IFA $\cap$ ELISA) is optimal if q > \$31.00. The PVP and 1-PVN values for the rules are given in Table 12, where $T_1$ = ELISA and $T_2$ = ELISA.

Table 12

| Rule | PVP | 1-PVN |
|---|---|---|
| $T_1$ | .92 | .007 |
| $T_2$ | .65 | .002 |
| $T_1 \cap T_2$ | .94 | .008 |
| $T_1 \cup T_2$ | .67 | .0003 |

## 6. Remarks

We have attempted to present the ingredients that are required in designing a mass screening program, as well as how to conduct one in an optimal manner. Although we have restricted our concern to at most two binary tests, it is conceptually clear how one would proceed formulaically with several such binary tests. There also should be no great difficulty in managing such a program when several dichotomous covariates are also taken into account. Greater complexity arises when including several continuous covariates into such a decision

theoretic framework.

7. Acknowledgement

We are indebted to J.L. Gastwirth for helpful suggestions regarding the data. The work was sponsored in part by an NIH grant GMS-25271.

## References

Burkhardt, U., Mertens, Th., and Eggers, H.J. (1987). "Comparison of two commercially available anti-HIV ELISAs: Abbott HTLV III EIA and DuPont HTLV III - ELISA." Journal of Medical Virology, 23:217-224.

Evans, R.P., Shanson, D.C. and Mortimer, P.P. (1987). "Clinical evaluation of Abbott and Wellcome enzyme linked immunosorbent assays for detection of serum antibodies to human immunodeficiency virus (HIV)." Journal of Clinical Pathology, 40:552-555.

Gastwirth, J.L. (1987). "The statistical precision of medical screening tests." Statistical Science, 2:213-38.

Gastwirth, J.L. and Hammick, P. (1989). "Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors." Journal of Statistical Planning and Inference, 22:15-27.

Gastwirth, Johnson and Reneau (1988). "Bayesian analysis of screening data: Application to AIDS in blood donors." Preprint.

Gatsonis, C. and Iyengar, S. (1987). "Bayesian inference for screening tests." Preprint.

Geisser, S. (1987). "Comment on 'The statistical precision of medical screening tests', by J.L. Gastwirth." Statistical Science, 2:231-2.

Ghi, V., De Palma, M., Vecchi, C., Pellegrino, F., Mongiardo, N. Cocchi, I., De Rienzo, B. and Squadrini, F. (1988). "Comparison between enzyme-linked immunosorbent and IF assays in the serological diagnosis of HIV infections." Microbiologica, 11:81-84.

Hirsch, V.M., Searcy, G.P., Bellamy, J.E.C. (1982). "Comparison of ELISA and immunofluorescence assays of feline leukemia virus antigens in blood of cats." Journal of the American Animal Hospital Association, 18:933-938.

Jarrett, O., Goldberg, M.C., Wejser, K. (1982). "A comparison of three methods of feline leukemia virus diagnosis." Veterinary Record, 110:325-328.

Johnson and Gastwirth (1988). "Bayesian inference for medical screening tests: Approximations useful for the analysis of AIDS data." Preprint.

Kurlander, A.B., Hill, B.H., Enterline, P.E. (1955). "An evluation of some commonly used screening tests for heart disease and hypertension." Journal of Chronic Disease, 2:427-429.

Lundberg, C. (1988). "Serological diagnosis of human immunodeficiency virus infection by Western Blot testing." Journal of the American Medical Association, 260:674-679.

Nusbacher, J., Chiavetta, J., Naiman, R., Buchner, B., Scalia, V. and Horst, R. (1986). "Evaluation of a confidential method of excluding blood donors exposed to human immunodeficiency virus." Transfusion, 26:539-541.

Sandstrom, E.G., Schooley, R.T., Ho, D.D., Byington, R., Sarnyadharan, M.G., MacLane, M.E., Essex, M., Gallo, R.C., and Hirsch, M.S. (1985). "Detection of human anti-HTLV-III antibodies by indirect immunofluorescence using fixed cells." Transfusion, 25:308-312.

Tierney, L. and Kadane, J. (1986). "Accurate approximations for posterior moments and marginal densities." Journal of the American Statistical Association, 81:82-6.

Figure 1

REGIONS OF OPTIMALITY